13. Linking Core Wave, Topical Module, and Longitudinal Research Files

In many situations, a single Survey of Income and Program Participation (SIPP) data file will not contain the information needed for a project. Because only limited core information is included on the topical module files, analysts often need to merge data from the core wave or longitudinal research files with topical module information. Also, they may need to link two or more topical module files, each containing data on a different topic and collected in different waves. And there are situations in which it is necessary to merge data from the core wave files with data from the longitudinal research files. Those situations arise because not all of the core wave content is included on the longitudinal research files (e.g., calendar month weights are only on the core wave files).[1] This chapter describes procedures for linking core wave, topical module, and full panel data files.

This chapter assumes a working knowledge of the files that will be linked.[2] Analysts who are not familiar with those files should read the following before proceeding with this chapter:

- Chapter 9 for an overview of the SIPP data files;
- Chapter 10 for a discussion of the core wave files;
- Chapter 11 for a discussion of the topical module files; and
- Chapter 12 for a discussion of the longitudinal research files.

In all cases, this chapter describes procedures for linking person records across files. It does not discuss procedures for linking households or families because those procedures become problematic when working with longitudinal data.[3]

[1] Even when the same variables are on both the core wave and longitudinal research files, the data may not be the same. Different edit and imputation procedures are used for these two types of files. Prior to the 1996 Panel, all edit and imputation procedures applied to the core wave files worked entirely *within* the given file. Information from previous waves or later waves was not used. Beginning with the 1996 Panel, edit and imputation procedures applied to the core wave files make greater use of information from previous waves. However, because the core wave files are processed as the data become available, it is not possible to make use of information from future waves. The edit and imputation procedures applied to the longitudinal research files, however, make use of each person's full longitudinal record. There are many times when the preferred data for a study will be on the longitudinal research files but the weights will be on the core wave files.

[2] This chapter does not discuss the longitudinal research file from the 1996 Panel because, as of this writing, it is not available. That information will be added to an updated version of this chapter once the file becomes available. In the interim, the only information included in this chapter on the 1996 longitudinal research file is the new variable names being used in the 1996 Panel data files.

[3] Difficulties arise when unit composition changes over time. In those situations, there is no unambiguous way to define longitudinal households and families, and many *ad hoc* procedures run the risk of introducing biases into

This chapter begins with a discussion of the mechanics involved in linking SIPP data files. The procedures are straightforward and easily implemented. In each case there are three basic steps:

1. Create data extracts from each of the files to be linked;
2. Sort the files in common order by using the variables identified as match keys; and
3. Merge the files.

There are two general formats that the final files can take. This chapter refers to these as person-month format (the format of the current core wave files) and person-record format (the format of the longitudinal research files).[4] The choice of format will be a function of the planned analysis and the software that will be used for that analysis. Where appropriate, procedures for generating each type of data file are described.

After discussing the mechanics of linking SIPP files, this chapter discusses why nonmatches occur and suggests ways to deal with them.

For the 1996 Panel, most variable names changed from those of previous panels. To aid users working with pre-1996 panel files, this chapter presents both the old and the new variable names when the text applies to both. In the main body of the text, the old names are presented in parentheses following the new names. For example, the sample unit ID variable name, which is SSUID in the 1996 Panel, was SUID in previous panels; it is written in this chapter as SSUID (SUID). In tables, a variety of methods are used to present both the old and the new names.

Procedures for Linking Files

There are six types of merges that SIPP users commonly need to perform:

1. Person-month records within a core wave file can be linked, creating a single wide record for each person rather than a record for each person for each month;[5]
2. Two or more core wave files can be linked together;
3. Core wave files can be linked to longitudional research files;

analyses of those units. The alternative approach that has gained acceptance in the research community involves assigning to people the characteristics of the households or families to which they belong at each point in time. Subjects can then be followed over time, as can the characteristics of the households or families to which they belong. One exception to the longitudinal household problem is with program units (e.g., food stamp units), where program rules can be used to define when changing composition constitutes the formation of a new unit (as opposed to changed composition of an existing unit). For discussions of the issues involved in studying longitudinal households and families, see McMillen and Herriot (1985), Duncan and Hill (1985), Citro et al. (1986), and Kalton et al. (1987).

[4] Some software (e.g., Stata) refers to this as "wide" format, while the person-month format is referred to as "long."

[5] This procedure transforms the current format of the core wave files into a format similar to that used prior to the 1990 Panel, a format analogous to that used for the longitudinal research files.

4. Two or more topical module files can be linked to each other;
5. Topical module files can be linked to core wave files; and
6. Topical module files can be linked to longitudinal research files.

This chapter addresses each of these merges in turn.

Linking Within a Core Wave File—Transforming the Person-Month Format into the Person-Record Format

This procedure transforms the person-month-format core wave files (with one record per person per month) into a single wide record per person (the format used for the core wave files before the 1990 Panel). As well as being useful in its own right, reformatting is often a necessary first step when merging core wave files with data from either the topical module files or from the longitudinal research files.

Two approaches for this link are described. Programmers using third-generation languages, such as FORTRAN and PL/1, typically use the first approach. Programmers using fourth-generation languages, such as SAS and SPSS, typically use the second approach.

The first approach (using FORTRAN) contains four steps:

1. Sort the file by person and reference month, using the following variables: sample unit ID [SSUID (SUID)], entry address ID [EENTAID (ENTRY)], person number [EPPPNUM (PNUM)], and reference month [SREFMON (REFMTH)].[6] This is the sort order the Census Bureau uses for the core wave files. If the file being used is in its original sort order, this step can be skipped.
2. Define and initialize monthly variable arrays to some "missing data" code. Users should be careful to choose initial values outside the range of legal values for the variables of interest. For example, the variable TAGE (AGE) would be defined as an array of four elements, and each element could be initialized to –9 (an age that no one can have); the variable TPTOTINC (TOTINC) would be defined as an array of four elements and each element could be initialized to –999999 (a negative value outside the range of the variable), and so on.
3. Read each person's corresponding person-month record and put the information into the appropriate element of the array.
4. Write the person-based record from the information stored in the arrays.

The second approach (using SAS) also contains four steps:[7]

[6] In the 1996 Panel, the entry address is no longer needed to uniquely identify people. Its continued use will not create any problems; it is simply redundant information for purposes of identifying SIPP sample members.

[7] An alternative procedure that may be useful in many cases uses SAS Proc Transpose. Stata also has a procedure—reshape—that can accomplish this task.

1. Sort the file by person and reference month, using the following variables: sample unit ID [SSUID (SUID)], entry address ID [EENTAID (ENTRY)], person number [EPPPNUM (PNUM)], and reference month [SREFMON (REFMTH)]. This is the sort order used by the Census Bureau for the core wave files. If the file being used is in its original sort order, this step can be skipped.
2. Write out four files, each one containing the person ID variables and the variables for 1 of the 4 months. For example, file1 would have the person ID variables [SSUID (SUID), EENTAID (ENTRY), and EPPPNUM (PNUM)] and the variables for month one, file2 would have the person ID variables and the variables for month two, and so on.
3. Rename the (monthly) variables in each of the four files to unique names. For example, the variable names in file1 might be TAGE1 (AGE1) and PTOTINC1[8] (TOTINC1); in file2 the variable names might be TAGE2 (AGE2) and PTOTINC2 (TOTINC2).
4. Merge the four files together, using SSUID (SUID), EENTAID (ENTRY), and EPPPNUM (PNUM) as the match keys.

The SAS code in Figure 13-1 performs the above steps.

The person-month format of the core wave files (before reformatting) is illustrated in Table 13-1. Person number 101 is in the sample all 4 months, person number 102 is in the sample all 4 months, person number 201 is in the sample for 2 months, and person number 202 is in the sample for 1 month. The person-record format (after reformatting) is illustrated in Table 13-2. Missing data are indicated by a single period, the default missing data code in SAS. For the FORTRAN example, the missing data would have codes of –9 and –999999.

Linking Two or More Core Wave Files

There are three reasons to link two or more core wave files:

1. To create an analysis file for one or more calendar months containing data from all four rotation groups. For example, data for March 1994 are contained in the Wave 7 file (of the 1992 Panel) for rotation groups 4 and 1, and in the Wave 8 file for rotation groups 2 and 3. (Data for the same calendar month are also in Waves 4 and 5 of the 1993 Panel.)
2. To create an analysis file containing more than 4 months of information for each person. This linkage is of primary interest to users of the 1996 Panel, beause longitudinal research files for all other panels are available from the Census Bureau.
3. As preparation for merging core wave data with data from either the topical module files or the longitudinal research files.

[8] Because variable names in SAS are limited to eight characters, the monthly variable name is shortened from TPTOTINC1 (nine characters) to PTOTINC1 (eight characters).

Figure 13-1. Sample SAS Code to Change the Core Wave Files from Person-Month Format to Person-Record Format from Wave 2 of the 1996 Panel

```
/*
   this creates the initial extract from the full core wave file
*/
data allmnths;
   set corewv962
      (keep =
         ssuid
         eentaid
         epppnum
         srefmth
         tage
         tptotinc
      );
run;

/*
   sort the data - if the master file was in its original order, this
   step is not needed
*/
proc sort;
   by ssuid eentaid epppnum srefmth;
run;

/*
   write out 1 file for each of the four months, renaming variables in
   the process
*/
data
   file1
      (rename =
         (tage = tage1
          tptotinc = ptotinc1
          srefmth = srefmth1
         )
      )
   file2
      (rename =
         (tage = tage2
          tptotinc = ptotinc2
          srefmth = srefmth2
         )
      )
   file3
      (rename =
         (tage = tage3
          tptotinc = ptotinc3
          srefmth = srefmth3
         )
      )
```

(figure continues)

When text copy applies to both 1996 and pre-1996 panel files, pre-1996 variable names appear in parentheses following 1996 variable names.

Figure 13-1. Sample SAS Code to Change the Core Wave Files from Person-Month Format to Person-Record Format from Wave 2 of the 1996 Panel (*continued*)

```
   file4
      (rename =
         (tage = tage4
          tptotinc = ptotinc4
          srefmth = srefmth4
         )
      )
   ;

   set allmnths;

   select (srefmth);
      when (1) output file1;
      when (2) output file2;
      when (3) output file3;
      when (4) output file4;
   end;
run;

/*
   merge the 4 "monthly" files together, forming the final file
*/
data newfile;
   merge
      file1
      file2
      file3
      file4
      ;
   by ssuid eentaid epppnum;
run;
```

Creating files in the person-month format is straightforward. In this instance, the files from each of the contributing core wave files simply need to be sorted and interleaved to create the final analysis file. The final sort order would likely be based on SSUID (SUID), EENTAID (ENTRY), EPPPNUM (PNUM), SWAVE (WAVE), and SREFMON (REFMTH).

If a person-record format (with just one record per person) is desired, the first step is interleaving the files to create the person-month-format file. Then, using that as the input file, analysts can apply the procedures described in the preceding section to generate a file with a single wide record for each person. There will be up to 4 months of data for each wave used. In the example from Tables 13-1 and 13-2, if three waves of data are being combined, the final file will have 12 values for SREFMON (REFMTH), TAGE (AGE), and TPTOTINC (TOTINC). In the SAS program code, the names would likely be REFMTH1–REFMTH12, TAGE1–TAGE12, and TOTINC1–TOTINC12.

Users attempting to create their own longitudinal databases from the core wave files should proceed cautiously. The edit and imputation procedures applied to the core wave files for the

When text copy applies to both 1996 and pre-1996 panel files, pre-1996 variable names appear in parentheses following 1996 variable names.

Table 13-1. Example of the Core Wave Person-Month File Structure

Sample Unit ID [SSUID (SUID)]	Entry Address ID [(EENTAID (ENTRY)]	Person Number [EPPPNUM (PNUM)]	Reference Month [(SREFMON (REFMTH)]	Age [TAGE (AGE)]	Total Income [(TPTOTINC (TOTINC)]
123456781000	011 (11)	0101 (101)	1	42	$2000
123456781000	011 (11)	0101 (101)	2	42	$2100
123456781000	011 (11)	0101 (101)	3	42	$2000
123456781000	011 (11)	0101 (101)	4	43	$2000
123456781000	011 (11)	0102 (102)	1	41	$ 500
123456781000	011 (11)	0102 (102)	2	41	$ 500
123456781000	011 (11)	0102 (102)	3	41	$ 0
123456781000	011 (11)	0102 (102)	4	41	$ 0
123456781000	011 (11)	0201 (201)	2	18	$ 200
123456781000	011 (11)	0201 (201)	3	18	$ 200
123456781000	011 (11)	0201 (201)	4	18	$ 200
123456781000	011 (11)	0202 (202)	2	2	$ 0
123456781000	011 (11)	0202 (202)	3	2	$ 0
123456781000	011 (11)	0202 (202)	4	2	$ 0

Table 13-2. Example of the Core-Wave Wide-Record/Person File Structure (After Applying the Program in Figure 13-1 to the Data in Table 13-1)

Sample Unit ID [SSUID (SUID)]	Entry Address ID [EENTAID (ENTRY)]	Person Number [EPPPNUM (PNUM)]	Reference Month (SREFMTH)[a]				Age (TAGE)[b]				Total Income (PTOTINC)[c]			
			1	2	3	4	1	2	3	4	1	2	3	4
123456781000	011 (11)	0101 (101)	1	2	3	4	42	42	42	43	$ 2000	$ 2100	$ 2000	$ 2000
123456781000	011 (11)	0102 (102)	1	2	3	4	41	41	41	41	$ 500	$ 500	$ 0	$ 0
123456781000	011 (11)	0201 (201)	.	2	3	4	.	18	18	18	.	$ 200	$ 200	$ 200
123456781000	011 (11)	0202 (202)	.	2	3	4	.	2	2	2	.	$ 0	$ 0	$ 0

Note: . = missing.

[a] 1 = SREFMTH1, 2 = SREFMTH2, 3 = SREFMTH3, 4 = SREFMTH4.

[b] 1 = TAGE1, 2 = TAGE2, 3 = TAGE3, 4 = TAGE4.

[c] 1 = PTOTINC1, 2 = PTOTINC2, 3 = PTOTINC3, 4 = PTOTINC4.

SIPP panels prior to the 1996 Panel were all "within wave" procedures. This means that the edits and imputations applied to a person's records in one wave were independent of those in other waves. Imputation procedures for most of the core wave files from the 1996 Panel are different. The new procedures do make use of information from the preceding wave. When linking data across waves, apparent changes in income, program participation, labor force behavior, or most other outcomes could be due to real changes reported by the respondent, or they could be an artifact of the data editing and imputation performed by the Census Bureau. Although this problem arises primarily with the core wave files from panels prior to 1996, it is also true of the 1996 Panel.[9]

[9] The new imputation procedures for the 1996 Panel are expected to introduce less error than procedures used for earlier panels. Thus, the number and magnitude of spurious changes (as well as falsely imputed stability) should be reduced. Even so, imputation errors will occur, and caution is advised when using the core wave files for longitudinal research.

There are two ways to identify cases with edited or imputed data. In panels prior to 1996, the entire record was imputed if (1) MIS5 = 2 and MISj = 1 for j = 1, 2, 3, or 4 or (2) INTVW = 3 or 4. The record was imputed in the 1996 Panel if EPPINTVW = 3 or 4. In the 1996 Panel, persons with Type Z noninterviews with prior wave information have their items imputed with procedures that use their prior wave responses. The relatively few cases with no prior wave information (those in Wave 1 and those in Waves 2–12 who are new to the sample) have their records imputed with the Type Z procedure used in the pre-1996 files. For all panels, if the record was not imputed, it is necessary to check the allocation (imputation) flags associated with the variables of interest. Once identified, users might need to implement some form of longitudinal editing and imputation or distinguish in their analyses between "real" changes and those that may result from the core wave data processing procedures.

Basic demographic information, such as age, race, and sex, can also appear to change from one wave to the next. In these instances, changes reflect corrections made in later interviews to information collected in earlier interviews; it is generally safe to assume the most recent data are correct.

When using the core wave files for longitudinal research, analysts should also note that the sample weights included on the core wave files are calendar month specific. These weights may not be appropriate for the planned longitudinal analyses. Chapter 8 has a detailed discussion of how to use the sample weights provided with the SIPP files.

Linking Core Wave Files to Longitudinal Research Files

There are relatively few circumstances in which the core wave and full panels files need to be linked because, for the most part, they contain the same information.[10] In general, if the same information is available from both the core wave and longitudinal research files, the information from the longitudinal research files is preferable because the edit and imputation procedures used for the longitudinal research files are believed to introduce less error than the procedures used for the core wave files.[11] However, some core information is contained only on the core wave files, and, therefore, at times it will be necessary to merge the core wave and longitudinal research files.

The following steps are necessary to link data from the core wave files with data from the full panel files:

1. Create data extracts from the core wave and longitudinal research files;
2. Put the two extracts into the same format (either person-month format or person-record format);

[10] Because the 1996 longitudinal research file is not complete yet, the discussion in this section pertains only to files for earlier panels. A revised version of this chapter will be available on the Census Bureau SIPP Web site (http://www.sipp.census.gov/sipp/) when the 1996 longitudinal research file is completed.

[11] See footnote 1.

3. Sort the extracts into the same order; and

4. Merge the extracts, creating the final file.

The variables that uniquely identify people in the core wave and longitudinal research files have different names. Table 13-3 shows the names for the three variables needed to match people across those files for panels prior to 1996.[12]

Table 13-3. Variables Identifying People in the Core Wave and Longitudinal Research Files for Panels Prior to 1996

Variable	Core Wave Files		Longitudinal Research Files
Sample Unit ID	SUID	is matched to	PP-ID
Entry Address ID	ENTRY	is matched to	PP-ENTRY
Person Number	PNUM	is matched to	PP-PNUM

If the final file will be in person-record format, these are the only variables needed for the sort and merge operations (steps 3 and 4, above). If the final file will be in person-month format, then WAVE and REFMTH are also needed.

Figure 13-2 shows the SAS code to transform data from the longitudinal research files in wide-record format into the person-month format used in the core wave files. The program creates a person-month format file from the 1993 longitudinal research file.

Because SAS does not allow variable names with embedded dashes, the "-" characters in the variable names have been replaced with underscore ("_") characters. The 1993 Panel had 10 waves, so the output file will have up to 40 monthly records for each person: no records are written for any months when pp_mis is not equal to 1. The program creates a data set with seven variables: SUID (renamed from PP_ID), ENTRY (renamed from PP_ENTRY), PNUM (renamed from PP_PNUM), REFMTH (which ranges from 1 to 4), WAVE (which ranges from 1 to 10), AGE, and TOTINC.

The REFMTH variable is computed as modulus (i/4) if it is not equal to 0, or 4 if is equal to 0. The modulus is the remainder from the division, so in month six of the panel the quantity is modulus (6/4) = 2, in month seven it is modulus (7/4) = 3, and in month eight it is 4 (since the remainder from the division of 8 by 4 is 0).

The wave is computed as the first integer greater than or equal to i/4. For month one, i/4 = 0.25, so wave = 1. For month four, i/4 = 1, so wave = 1. For month 17, 17/4 = 4.25, so wave = 5.

The file created by the program in Figure 13-2 could be merged with an extract from the core wave files from the 1993 Panel, using SUID, ENTRY, PNUM, WAVE, and REFMTH as the match keys. If the longitudinal research file was in its original sort order, the file created by the program in Figure 13-2 will already be sorted by this set of match keys.

[12] Current plans call for using consistent variable names across all files from the 1996 Panel.

Figure 13-2. Sample SAS Code to Change the Longitudinal Research Files from Person-Record Format to Person-Month Format for Panels Prior to 1996

```
Data pmonth
   (keep =
      pp_id
      pp_entry
      pp_pnum
      refmth
      wave
      age
      totinc
    rename =
      (pp_id = suid
       pp_entry = entry
       pp_pnum = pnum
      )
   );

   /*
      this example works with the 1993 SIPP panel - 10 waves
   */
   set sipp93fp
      (keep =
         pp_id
         pp_entry
         pp_pnum
         pp_mis1 - pp_mis40
         age1 - age40
         totinc1 - totinc40
      );

   /*
      define arrays to ease the programming burden
   */
   array ages {40} age1 - age40;
   array totincs {40} totinc1 - totinc40;
   array pp_mis {40} pp_mis1 - pp_mis40;

   do i = 1 to 40;                  /* for each month */
      if (pp_mis{i} eq 1) then do;  /* if pp_mis is 1, use the data */
         age = ages{i};             /* the age in this month */
         totinc = totincs{i};       /* total income this month */

         j = mod(i,4);
         if (j eq 0) then refmth = 4;/* the reference month */
         else refmth = j;

         wave = ceil(i/4);          /* the wave */
         output;                    /* write out the record */
      end;
   end;
run;
```

When text copy applies to both 1996 and pre-1996 panel files, pre-1996 variable names appear in parentheses following 1996 variable names.

Values for AGE and TOTINC from the core wave and longitudinal research files will not match for all people in all months because the core wave files and the longitudinal research files are subjected to different edit and imputation procedures.

In addition, beginning with the 1991 Panel, a missing wave imputation procedure has been applied to the longitudinal research files: people who had missing data from one wave but complete data from the two adjacent waves had data imputed for the missing wave in the longitudinal research files.[13] This means that some people will have data in the longitudinal research files for months in which they have no records in the associated core wave files (those who were not Type Z nonrespondents).

Linking Two or More Topical Module Files

At times it will be necessary to merge data from two or more topical module files. Any project that studies the relationship between subject areas covered by different topical modules will require such a merge. One example might be a study of the relationship between the use of health care services (collected in Wave 3 of the 1993 Panel) and medical expenses (collected in Wave 4 of the 1993 Panel).

The mechanical process of linking topical module files is relatively straightforward. The topical module files all have the same format (one record per person) and variable names, for the ID variables are consistent across the topical module files: individuals are uniquely identified by the combination of SSUID (ID), EENTAID (ENTRY), and EPPPNUM (PNUM).

However, a number of cautions should be noted:

1. Prior to the 1996 Panel, there were instances in which the same variable name was used in different topical module files for different variables. For example, in the 1990 Panel, TM8400 was used in the Wave 2 topical module for a variable that indicates whether the respondent completed 12th grade. The same variable name was used in the Wave 6 topical module to indicate whether the respondent was a parent of children under 21 years of age living in his or her household.
2. Not all people with records in one topical module file will have records in another topical module file. In the topical module files from the 1996 Panel, there will generally be a record for each person who was a responding SIPP household member in the fourth month of the wave's core reference period. Prior to the 1996 Panel, all household members in the interview month have topical module records for a given wave. However, household composition changes from one wave to the next: some people leave SIPP households and others join SIPP

[13] Many of these situations arise with Type Z nonrespondents: nonresponding people who live in households with other responding sample members. Type Z nonrespondents in the pre-1996 core wave files and those in the 1996 Panel files with no prior wave information were subjected to a whole-record imputation procedure, described in Chapter 10. These people would have records in the core wave files, but different information—because it was imputed using different procedures—in the longitudinal research files.

households, and this changing composition is reflected in the topical module files. Also, in the 1996 Panel, some people who were nonrespondents in month four of one wave may have been respondents in month four of another wave. Thus, when topical module files are merged, there will be a nontrivial number of nonmatches: people with data from only one of the topical modules. Nonmatches are addressed in greater detail later in this chapter.

3. Choosing appropriate weights is complicated by the fact that there are a substantial number of nonmatches across topical modules. One solution is to use one of the weights from the longitudinal research files. Chapter 8 gives a detailed discussion of the SIPP weights.

Often it will be necessary to merge additional information (such as sample weights) from the core wave or longitudinal research files when working with multiple topical modules.

Users interested in measuring change with data from the topical module files (such as changes in asset holdings, or changes in health or disability status) should proceed with caution. First, in some instances measurement error is large relative to the actual changes that have taken place. One example is found in the topical modules that measure levels of household assets and liabilities.[14] Although the topical modules can provide estimates of aggregate-level changes in those instances, users should not attempt to measure those changes at the individual level. Also, the edit and imputation procedures applied to the topical module files are all "within wave" procedures. This means that the edits and imputations applied to a person's records in one wave are independent of those in other waves. When data are linked across waves, apparent changes could be due to real changes reported by the respondent or they could be artifacts of the data editing and imputation performed by the Census Bureau.

There are two ways to identify cases with edited or imputed data. In panels prior to 1996, the entire record was imputed if (1) PP-MIS5 = 2 and PP-MISj = 1 for j = 1, 2, 3, or 4 or (2) INTVW = 3 or 4. In the 1996 Panel, the record was imputed if (1) EPPMIS4 = 2 or (2) EPPINTVW = 3 or 4. In the 1996 Panel, persons with Type Z noninterviews who have prior wave information have their records imputed with procedures that use their prior wave responses. For persons with no prior wave information (those in Wave 1 and those in Waves 2–12 who are new to the sample), the Type Z imputation procedure is used. On all panels, users should check the imputation flags associated with the variables of interest.

Linking Topical Module Files to Core Wave Files

Because the topical module files contain only limited information from the SIPP core, there will be many times when it is necessary to merge data from the topical module files with data from the SIPP core. One source of these data is the core wave files.[15]

[14] See the *SIPP Quality Profile*, 3rd Ed. (U.S. Census Bureau, 1998a) and SIPP Working Paper series for discussions of this issue as it relates to this and other SIPP topical modules.

[15] The next section describes procedures for merging topical module files with data from the longitudinal research files.

The first decision that must be made is which core wave file to use. Special attention should be paid to the reference periods for the topical module items of interest. In the 1996 Panel, topical module questions refer to either month four of the wave's core reference period, or to a longer period in the past (such as the preceding 12 months or the prior calendar year). In those instances, information would come from the month-four records of the core wave files from the same wave (and possibly from earlier months and waves). Prior to the 1996 Panel, many topical module items referred to conditions in the interview month. The interview month, however, is not included as a separate record in the core wave file for the same wave as the topical module.[16] Rather, core information for the interview month of one wave is found in the month-one information from the following wave. For example, the interview month for Wave 3 is month 13 in the SIPP panel, and core data for month 13 are collected as the first reference month of Wave 4.[17] Commonly used reference periods for topical module items are the current (interview) month (month one of the *next* wave), the previous month (month four of the current wave), the previous 4 months (the full reference period for the current wave), and the previous year.

The topical module files have one record per person, while the core wave files have up to four records for each person (one record per person for each month the person was a SIPP sample member). There are at least three options available when merging topical modules with data from the SIPP core wave files:[18]

1. Pick a single month from the core wave files. For example, if the topical module items use the interview month as their reference period, it may make sense to use records for month one from the core wave files from the *next* wave.
2. Spread the topical module data across all records from the core wave file. That results in a final file in person-month format.
3. Create a single record for each person from the appropriate core wave file and merge the topical module data to that record. This results in a final file in the person-record format with the same monthly detail as in the second option described above.

The steps involved are as follows:

1. Create an extract from the core wave file(s) of interest.
2. If a single record for each person is desired, apply the algorithm in Figure 13-1, which is described in the section entitled Linking Within a Core Wave File—Transforming the Person-Month Format into the Person-Record Format.

[16] Some of the interview month information is contained on the records for the four reference months of the wave. But in the person-month-format file there is no separate record for the interview month itself.

[17] Information collected during the interview month of one wave may not match the information collected about the same calendar month in the subsequent wave. In the 1996 Panel, dependent interviewing techniques and other checks made possible with CAI are used to help resolve those inconsistencies.

[18] Yet another option is to create a single record from the core wave files containing aggregate measures for the reference period of interest. For example, it might make sense to create a single record from the "current" core wave file with total income received during all 4 months of the wave's reference period. Or the average number of hours worked per week during the previous 4 months might be appropriate. Once the aggregate record is created, the merge step is similar to the others described in this section.

3. Sort the core wave extract using SSUID (SUID), EENTAID (ENTRY), and EPPPNUM (PNUM) as the sort keys. These three variables uniquely identify people in the core wave files. If the core wave extract is in the person-month format, include SREFMON (REFMTH) as the final sort key.
4. Create an extract from the topical module file of interest. Sort the topical module extract using SSUID (ID), EENTAID (ENTRY), and EPPPNUM (PNUM) as the sort keys.
5. For the 1996 Panel, merge the core wave extract with the topical module extract; use SSUID, ENTAID, and EPPPNUM as the sort keys. For panels prior to 1996, merge the core wave extract with the topical module extract; use the sort keys shown in Table 13-4.

Table 13-4. Variables Identifying People in the Topical Module and Core Wave Files for Panels Prior to 1996

Variable	Topical Module Files		Core Wave Files
Sample Unit ID	ID	is matched to	SUID
Entry Address ID	ENTRY	is matched to	ENTRY
Person Number	PNUM	is matched to	PNUM

When data from panels prior to 1996 are used, there will likely be a nontrivial number of nonmatches between the core wave files and the topical module files. That will be true even when a topical module is merged with core data from the same wave, because people who were members of a SIPP household in the interview month but not during the previous 4 months will have records in the topical module files but not in the core wave files.

Linking Topical Module Files to Longitudinal Research Files from Pre-1996 Panels

While topical module files can be linked with data from the core wave files, there are many times when it will be necessary or desirable to use the longitudinal research files instead.[19] For example, if the full panel weights[20] are needed for the planned analysis, they must come from the longitudinal research files. When the same core items are available from the core wave and the longitudinal research files, analysts may prefer to use the longitudinal research files because the edit and imputation procedures used for them are believed to introduce less error than the procedures used for the core wave files.

[19] Because the full panel longitudinal research file for the 1996 SIPP was still under development at the time this chapter was written, it is not yet possible to describe procedures for using that file. A revised version of this chapter will be available once the longitudinal research file for the 1996 Panel is released to the public.

[20] Chapter 8 discusses the SIPP weights, their derivation, and use.

The steps involved are as follows:

1. Create an extract from the longitudinal research file.
2. If a file in the person-month format is desired, apply the algorithm described in the section above, Linking Core Wave Files to Longitudinal Research Files. The example in Figure 13-2 can be adapted to that purpose, but the ID variables would need to be renamed to match those used in the topical module files rather than in the core wave files (Table 13-5).
3. Sort the full panel extract; use PP-ID, PP-ENTRY, and PP-PNUM as the sort keys. These three variables uniquely identify people in the longitudinal research files. If the full panel extract is in the person-month format, include WAVE and REFMTH as the final sort keys.
4. Create an extract from the topical module file of interest. Sort the extract; use ID (the variable name for the sample unit ID in the topical module files), ENTRY, and PNUM as the sort keys.
5. Merge the core wave extract with the topical module extract based on the sort keys described here and shown in Table 13-5.

Table 13-5. Variables Identifying People in the Topical Module and Longitudinal Research Files Prior to the 1996 Panel

Variable	Topical Module Files		Longitudinal Research Files
Sample Unit ID	ID	is matched to	PP-ID
Entry Address ID	ENTRY	is matched to	PP-ENTRY
Person Number	PNUM	is matched to	PP-PNUM

Because the longitudinal research files contain a record for every person who was ever a member of a SIPP household, every person with a record in a topical module file should have a record in the longitudinal research file. However, analysts working with a person-month-format file containing records only for months when PP-MIS = 1 may find nonmatches.

Nonmatches When Merging Files

SIPP is designed to follow a group of people over an extended period of time. This group includes only those who were interviewed in the first wave of the panel and the children subsequently born to or adopted by them.[21] Over the course of the panel, these original sample members are followed and interviewed every 4 months. Secondary sample members, on the

[21] In the 1993 Panel all original sample members were followed no matter what their ages. In all other panels, only original sample members aged 15 years or older are followed when they move to new addresses. In all cases, however, the SIPP data files contain a record for all people, including children, who reside in a household with at least one original panel member present.

other hand, are part of the SIPP sample only for as long as they continue to reside with at least one original sample member. As long as they are part of the SIPP sample, the secondary sample members are interviewed and included in the SIPP data files.

The problem of nonmatches occurs only when users merge across waves for any types of files. There is no matching problem when the same or different types of files are merged within the same wave.

As shown in Table 13-6, there are a variety of reasons why a person may be in one SIPP data file but not in another. All but one of the reasons are associated with people entering and leaving the SIPP sample:[22]

1. The original sample person may have left the SIPP sample universe (e.g., died, moved abroad, moved into military barracks, or moved into an institution);
2. The original sample person may have left the sample but is still in the sample universe (sample attrition);
3. The original sample person may have just reentered the SIPP sample universe (after living abroad, etc.);
4. The person is a newborn (a special case of a person joining the sample universe);
5. The secondary sample member has just begun living with an original sample person;
6. The secondary sample member no longer lives with an original sample member;
7. The person had data for a "missing wave" imputed in the longitudinal research file and has no records in the core wave or topical module files for that wave; and
8. Prior to the 1996 Panel, the Census Bureau may have intentionally altered the identification information of the person, thereby making it difficult to find a match for this person (in rare situations referred to as *merged households*).

A person's reason for leaving the SIPP sample is identified in the core wave and longitudinal research files. In the former, the variable name is ULFTMAIN (REALFT). In the longitudinal research files, the name is REASLEFT, and it has a value for each wave rather than each month. Figure 13-3 shows the variable values and corresponding descriptions.

Procedures for dealing with nonmatches vary, depending largely on the reasons the person entered or left the SIPP sample. A number of common scenarios are presented below.

[22] The SIPP following rules are described in greater detail in Chapter 2.

Table 13-6. Reasons for Nonmatches

Reasons	File #1 (earlier time period)	File #2 (later time period)
People Exiting the Sample		
Original sample people left the SIPP sample universe (left the population of inference) Person died Moved abroad—left sample universe Moved into military barracks—left sample universe Moved into an institution—left sample universe	Present	Not present
Original sample person exited from the sample (still in the sample universe but no longer in the sample) Refused to be interviewed	Present	Not present
Secondary sample person no longer lives with an original sample member	Present	Not present
People Entering the Sample		
Newborn	Not present	Present
Original sample person returns to SIPP sample universe (returns to the population of inference) Moved from abroad—entered sample universe Moved from military barracks—entered sample universe Moved from an institution—entered sample universe	Not present	Present
Original sample member returns to sample Original sample member agrees to be interviewed and returns to sample	Not present	Present
Secondary sample person now lives with an original sample member	Not present	Present
Missing Wave Imputation in the Longitudinal Research File (Beginning with the 1991 Panel)		
Person has data in the longitudinal research file but no data in the corresponding wave in the core wave or topical module files.		
Merged Households—Special Case		
"Old" version of the ID information	Present	Not present
"New" version of the ID information	Not present	Present

Exiting or Entering the Population

There is a fundamental distinction between situations in which people leave the sample because they leave the SIPP sample universe and situations in which they leave the sample despite the fact that they are still part of that population. The SIPP sample universe (the population that the SIPP sample represents) is the noninstitutionalized, resident population of the United States. It includes both civilian and military people; it includes adults and children who reside in the United States and outside of institutions.

People who leave this population because they die, move abroad, or move into institutions exit the SIPP sample because they are no longer a part of the population that SIPP represents. *In general, when nonmatches occur because people have entered or exited the population represented by the SIPP sample, data should not be imputed and weights should not be adjusted for the period when these people are outside of that population.* From the perspective of SIPP, these people do not exist when they are outside of the population represented by the sample.

Figure 13-3. Data Dictionary Entries for Variables Identifying the Reason a Person Left the SIPP Sample

Wave 2, 1996 Panel Core Wave File

```
D ULFTMAIN    2    606
T PE: UNEDITED VARIABLE - Main reason left Household
     What is the main reason ... left the household?
U Movers from households which contain sample persons at the time of
   interview, movers from a household which splits into multiple
   households. Note: This is an unedited field and the universe is not
   exact.<BR>
V          0 .Not answered
V          1 .Deceased
V          2 .Institutionalized
V          3 .On active duty in the Armed Forces
V          4 .Moved outside of U.S.
V          5 .Separation or divorce
V          6 .Marriage
V          7 .Became employed/unemployed
V          8 .Due to job change - other
V          9 .Listed in error in prior wave
V         10 .Other
V         11 .Moved to type C household
```

1993 Full Panel Files

```
D REASLEFT    9    143  9   1
     Range = (0:9)
     Preedited reason for leaving the Household Control Card item 23
U Persons who left at any time during the reference period
  Subscript 1: not applicable for Observation 1
  Subscript 2 - 8: reason left in Observations 2 - 8
V   0 .Not applicable or not answered or nonmatch
V   1 .Left - deceased
V   2 .Left - institutionalized
V   3 .Left - living in armed forces barracks
V   4 .Left - moved outside of country
V   5 .Left - separation or divorce
V   6 .Left - person #201 or greater no longer living with sample person
V   7 .Left - other
V   8 .Entered merged household
V   9 .Interviewed in previous wave but not in sample
```

(figure continues)

When text copy applies to both 1996 and pre-1996 panel files, pre-1996 variable names appear in parentheses following 1996 variable names.

Figure 13-3. Data Dictionary Entries for Variables Identifying the Reason a Person Left the SIPP Sample (*continued*)

1993 Core Wave Files

```
D REALFT      2    521
     Reason for leaving the household
     Applicable when previous wave address ID is not equal to control
     card address ID
     Range=(00:00,05:12,25:31,99:99)
U All persons, including children, no longer in the household
V   00 .Not applicable or not answered
V   05 .Left - deceased
V   06 .Left - institutionalized
V   07 .Left - living in Armed Forces barracks
V   08 .Left - moved outside of country
V   09 .Left - separation or divorce
V   10 .Left - person #201+  no longer living with sample person
V   11 .Left - other
V   12 .Left - entered merged household
* Should have been deleted in a previous wave:
V   25 .Left - deceased
V   26 .Left - institutionalized
V   27 .Left - living in Armed Forces barracks
V   28 .Left - moved outside of country
V   29 .Left - separation or divorce
V   30 .Left - 201+ person no longer living with sample person
V   31 .Left - other
V   99 .Listed in error
```

The following examples help explain why weighting adjustments and imputation are problematic in these situations:

- *A person is in the SIPP sample at Time 1 but dies before Time 2.* In this case, the person is not part of the population at Time 2. In computing the aggregate (total) income of the population at Time 1, this person's income would be included. To impute income to this person for the Time 2 observation, analysts would compute an aggregate income that is too high: The person had no income at Time 2, and so none should be imputed.[23] If this case is dropped from the analysis file and the weights are inflated for the remaining sample, the estimate of the total population at Time 2 would be too high. Because this person was not a part of the population at Time 2, the weights for the remaining sample members should not be inflated to represent this individual.

[23] If the person had been alive with income that she or he did not report to the Census Bureau, an estimate of his or her unreported income would be imputed to the individual. Failing to impute that unreported income would mean that the income received by a member of the population is not represented anywhere in the sample. That value would result in a sample estimate of aggregate income in the population that was lower than the actual value in the population.

- *A person is overseas at Time 1 but at Time 2 is living with an original sample member in the United States.* At Time 1, this person was not part of the population represented by the SIPP sample. Because this person was not a part of that population, the SIPP sample should not be adjusted in any way to represent this individual.

A number of strategies are possible for dealing with cases in which nonmatches result from people entering or leaving the population represented by the SIPP sample. One approach is to drop those people from the analysis sample entirely. No adjustment would be made to the weights of the remaining cases. However, the definition of the population represented by the remaining sample would change. The remaining sample represents the population that existed at both Time 1 and Time 2. It does not represent anyone who either entered or left the population.

That approach has the advantage of being simple to implement. It also results in a clearly defined population of inference. Caution is necessary, however, to the extent that people entering and leaving the population are systematically different from those who are present throughout the period being studied: the remaining sample *cannot* be used to draw inferences about this other part of the population. People entering and leaving prisons and nursing homes, for example, likely have very different income profiles than the population that remains outside of these institutions over the period under study.

If event-history models are used to analyze the data, another approach is possible.[24] With these models, exits from the population can be treated as competing outcomes. For example, in a study of unemployment dynamics, a competing risks model might allow for three possible outcomes: spells of unemployment can end because (1) a person becomes employed, (2) a person exits the labor force, or (3) a person exits the population.[25]

Exiting the Sample but Remaining in the Population (Sample Attrition)

Sample attrition occurs when people leave the SIPP sample but remain a part of the population represented by that sample. In these instances the remaining sample generally should be adjusted to represent the full population, including the part of the population represented by those who leave the sample.

There are several options for handling such cases:

- *Impute the missing data and proceed.* This option is appropriate for researchers familiar with the statistical literature on imputation for missing data. A full discussion of this topic is well beyond the scope of this manual. Analysts are cautioned, however, against using the common practice of "substituting the mean" for missing data. That practice can yield biased estimates

[24] For a description of these methods, see, for example, Tuma and Hannan (1984).

[25] In actual applications, more than three outcomes would likely be modeled. The determinants of entering a nursing home, for example, are likely quite different from the determinants of entering a prison.

of multivariate statistics (such as regression coefficients) and generally leads to downward-biased estimates of standard errors.

- *Drop cases with missing data, adjust (poststratify) the weights for the retained cases, and proceed.* This poststratification involves several steps.

  1. Tabulate the weighted number of cases by various socioeconomic categories before dropping any cases.
  2. Repeat the tabulation after dropping the nonmatches.
  3. Compute adjustment factors by dividing the weighted numbers from step 1 (before dropping any cases) by the weighted numbers from step 2 (after dropping cases).
  4. Create a new weight variable by multiplying the original weight variable by the appropriate poststratification factor computed in step 3.

This situation requires caution. A user who drops records may introduce selection biases because those in the retained sample may be more stable than those who leave. For example, the fact that a (former) sample member has left may be associated with other changes in that person's life, such as giving birth, getting married, or getting a new job. Because the person left the sample, it is not possible to know from the available data what changes actually did occur in each case. Also, when records are dropped, the procedures for computing standard errors as described in the source and accuracy statements provided with the data will no longer apply. The procedures described in Chapter 7 for the direct estimation of standard errors should, however, work without any modification. If the number of cases lacking complete information is small relative to the full analysis sample (the full sample with positive weights), the biases introduced by dropping those cases also are likely to be small and this procedure may be a viable alternative.

- *If the longitudinal research file is available, use a subset of the cases with complete data for which Census Bureau–provided weights are available and proceed.* At the extreme, this procedure entails retaining only cases with positive full panel weights and using those weights for any analyses performed.[26] This is a conservative approach, but one that is relatively easy to implement because the weights already exist, they have already been adjusted for the observed sample attrition, and the population of inference is clearly defined.
- *Use other missing data methods to provide estimates and their standard errors.* A full discussion of these methods is beyond the scope of this manual. The methods are designed to make use of all available information from the cases with complete data without (directly) imputing data to cases with incomplete information. Interested users can consult the literature on the E-M algorithm for one example of how this can be done.[27] Also, Skinner et al. (1989) discuss model-based approaches to the analysis of complex surveys with missing data.

[26] The calendar year weights on the longitudinal research files are also options worth exploring. Chapter 8 provides a detailed discussion of the SIPP sample weights, their derivation, and use.

[27] For example, see Little and Rubin (1987). Users should also note that some statistical packages (e.g., SPSS) have incorporated more sophisticated options for handling missing data than have generally been available in the past.

Missing Wave Imputation in the Longitudinal Research Files Prior to 1996

Beginning with the 1991 Panel, a missing wave imputation procedure has been applied to the longitudinal research files: persons who had missing data from one wave but complete data from the two adjacent waves had data imputed for the missing wave in the longitudinal research files.[28] Some of those cases are Type Z nonrespondents and will have records with different data in the core wave files.[29] Other people will have data in the longitudinal research files for months when they have no records in the associated core wave or topical module files.

The correct procedure for dealing with the resulting nonmatches depends on which weight variables will be used. If the weights are coming from the core wave or topical module files, observations from the longitudinal research files not present in the cross-sectional files should be dropped. That is because the weights on the core wave and topical module files are computed for the samples in those files, samples that do not include the people who have had that wave imputed in the longitudinal research files.

If the weights are coming from the longitudinal research file, then other procedures must be used to deal with the missing data from the core wave and topical module files. In those instances, the procedures described for dealing with sample attrition should be considered.

Merged Households in Panels Prior to 1996

Finally, nonmatches can occur when the Census Bureau changes the ID numbers for sample members.[30] Prior to the 1996 Panel, there were two very rare occasions when this happened. The first occurred when two separate sampling units, each containing original sample members, were merged together, perhaps because of a marriage. In this situation, the people in one of the sampling units retained their identification information, while the people in the other sampling unit had their identification information changed to agree with the retained set. The person numbers of the changed set were modified to be between 180 and 199.

The second instance occurred when a SIPP household split into two new households (in which each new household gained a new sample person), which later recombined. For example, a

[28] Imputed waves can be identified on the longitudinal research files by using the WAVFLG variable.

[29] The data are different because different imputation procedures are used.

[30] Because the Census Bureau is using new procedures in the 1996 Panel, merged households will not be an identifiable source of nonmatches when files from the 1996 Panel are merged. Rather, they will appear no different from other situations where people enter and leave the SIPP sample, such as through marriages, divorces, deaths, and sample attrition. For example, in the 1996 Panel, there will be no way to identify which (if any) of the people who appear to have entered the sample in Wave 3 were also sample members who appear to have left the sample following Wave 2. The "new" sample members will be given person numbers in the same range as others who enter the sample in Wave 3, and no previous wave information will be attached to them. The new procedures greatly simplify the handling of these rare cases for both the Census Bureau and outside data users.

married couple separated in Wave 3, each moving in with a sibling. Both siblings were assigned a person number of 301, because they entered the sample in Wave 3 at different addresses. If the husband and wife reunited in Wave 6, bringing the siblings with them, one sibling's person number was changed. In this case, one of the siblings would have a person number of 301 and the other would have a person number of 680 (or some number between 680 and 699 because the households recombined in Wave 6).

Different file types (i.e., core wave, topical, and full panel) keep track of the changed ID values differently. If the move occurred after the first month of a reference period, the core wave file contains two records for the person whose identification information changed. The first record contains the original identification information of the person before the move and identifies the person as having exited the sample at the time of the move. The second record contains the new identification information after the move and identifies the person as having entered the sample at the time of the move. When the move occurs at the start of a reference period, only the second record is retained in the core wave file. The topical module file, however, contains only the second record, no matter when the move took place. The longitudinal research file contains both records, no matter when the move took place.

The easiest way to find these people is to search the core wave file for people with a previous wave identified as present, that is, PWSUID > 0 or PWENTRY > 0 or PWPNUM > 0. Users then need to decide how they want to handle these special cases. There are several possibilities:

- Change the identification information used in the waves before the move to the new values seen in the wave(s) after the move, and then merge the records using these ID values. This option is useful when working primarily with the person's core wave data after the move.
- Change the identification information in the waves after the move to the original values, and then use those ID values to merge records. This option is useful when working primarily with the person's core wave data before the move.
- Duplicate the person's record, and use the initial identification information with one record and the new identification information with the other record; then merge those records. With this approach, the weights for the duplicated records will need to be adjusted so that the duplicated weights sum to the original (unduplicated) weights.
- Treat this person as two people: once as someone who exits the sample at the time of the move and once as someone who enters the sample at the time of the move. That is how these cases are treated in the longitudinal research files. The weighting implications of this approach depend on the planned analysis.